Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:24:33
REFERENCE NO: 283

Author Names & Affiliations

- Alexander Ropelewski - Carnegie Mellon University, Pittsburgh Supercomputing Center
- Arthur Wetzel - Carnegie Mellon University, Pittsburgh Supercomputing Center

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computer Science, Image Analysis, High Performance Computing

Title of Submission

Cyberinfrastructure to Support Microscopy Imaging

Abstract (maximum ~200 words).

Advancements in the field of microscopy and imaging have pushed the boundaries of what was once thought possible in many fields of research. New techniques coupled with the application of new technologies allow researchers to probe further and with greater accuracy to answer increasingly complex questions. While these new techniques provide for far greater specificity of observation and increased sensitivity in regard to both resolution and frequency, the amount of data generated is increasing to a point where conventional systems are unable to manage it.

The largest single specimen optical and electron microscopy data sets, now already in the 100-terabyte range, are rapidly reaching petabyte scales. At the current time, there is no practical way to analyze, mine, share or interact with these large image data sets. The development of a national, scalable cyberinfrastructure solution for such data sets is extremely important as in the future there is expected to be continuous and sustained growth in data scale produced by these instruments.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Advancements in the field of microscopy and imaging have pushed the boundaries of what was once thought possible in many fields of research. New techniques coupled with the application of new technologies allow researchers to probe further and with greater accuracy to answer increasingly complex questions. While these new techniques provide for far greater specificity of observation and increased sensitivity in regard to both resolution and frequency, the amount of data generated is increasing to a point where conventional systems are unable to manage it.

The largest single specimen optical and electron microscopy data sets, now already in the 100-terabyte range, are rapidly reaching petabyte scales. Even more numerous, data volumes on the order of 1-10 terabytes each provide even larger aggregate data requirements and, regardless of individual scale, must be shared, viewed and analyzed in relation to other specimens with varying experimental conditions.

Over the last 10 years, the implementation of integrated systems, capable of imaging live cells or animals has become increasingly common. More importantly and presciently in the last 3 years, there has been a complete revolution in the capabilities of light microscopy such that data throughput and storage needs are potentially 10 times what they were 5 years ago. This change was initiated by multiple new technology developments, though primarily driven by the development of low cost, low noise complementary metal oxide sensor (sCMOS) cameras which allow large (between 2048x 2048 and 2,500x2,500 pixel) images with very low noise (1.03 e-/pixel/sec) at 100 frames/second with 16 bit depth. These cameras, which can easily generate 1.25 gigabytes of data/second, have been integrated into multiple fast collection systems. Similarly, in electron microscopy, high-throughput systems also exceeded 1-gigapixel capture rates.

There is also a growth in very fast resonant scanning confocal (ribbon scanning) and multiphoton microscopes. These platforms generate very large data sets, generally in the terabyte range. Data can be immediately housed using very fast multi-terabyte local storage systems. However, as the instruments that collect these data are extremely expensive and typically in heavy use it is not reasonable to perform analysis on the collection machine so the data must then be moved and analyzed post-hoc.

Furthermore, In addition to the volume and resolution of data, the use of consistent genetic resources in many animal models ensures that image data is somewhat portable between experiments, but the structure of datasets and the impracticality of sharing large datasets has limited the actual use of data from unrelated experiments. Currently there is no existing cyberinfrastructure solution for image data sets of this scale to enable the community to access, share, and use cross-experiment data effectively.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

At the current time, there is no practical way to analyze, mine, share or interact with large image data sets. A cyberinfrastructure solution that would act as a central and consistent repository of data from imaging technologies is needed. Specifically, brain imaging is ripe for the creation of a common resource. This cyberinfrastructure would store volumetric data, along with essential information about the experiment, in a scalable manner. It would also allow researchers to access, interact with, and analyze the stored data. The storage of such data sets is expected to grow into the exabyte range and thus will need to deal with the capacities, latencies, bandwidths, and processing speeds required by such extreme data-driven projects. Networking infrastructure is particularly needed to connect the high-speed networking backbone to the lab instrumentation generating the data itself, as a prerequisite to deposit and share the data produced by the instrumentation.

To create a useful resource, the cyberinfrastructure storage will need to initially store hundreds of petabytes of data on rotating storage and would be expected to need to grow into the exabyte range. The cyberinfrastructure will need to have a high-degree of reliability and redundancy build into the system. Depending on the value of the image data stored, the best solution may require a geographically distributed system capable of moving petabytes of data on demand from one physically distributed system to another and yet all addressable within a common framework. The system also must provide rich access control to maintain enclave access to pre-release and research access to sensitive, restricted-access data.

In addition to storage and networking, the need to have computational and visualization capabilities directly integrated with the storage is essential. The computational capacity of the cyberinfrastructure will need to include large memory computational nodes with general purpose GPUs which can be used for in-RAM data processing operations of these large image data sets and allow data sets to be compared within a reasonable amount of time and file system impact.

Finally, there are hardware and software barriers to visualization that need to be addressed. For example, users of large data sets will need to visually extract the most interesting regions for further analysis. This simple task requires interactively viewing the data sets from multiple scales, viewpoints, and other interactive renderings, in ways that are not being met by current techniques (for example, X11). The

integration of server-side GPU's and video-rate interfaces may be one way to alleviate these problems in the future.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."